Easy Approaches to Establishing Validity in a Task-Based Teacher Performance Assessment System

Individual Paper Presented at the Annual Meeting of the American Association of Colleges of Teacher Education February 2005

William Steve Lang, Ph.D., wslang@tempest.coedu.usf.edu
Judy R. Wilkerson, Ph.D., wilkerso@tempest.coedu.usf.edu
University of South Florida St. Petersburg

The Problem Defined

NCATE (2002) requires the measurement of knowledge, skills, and dispositions as part of its accreditation requirements for teacher education programs (Standard 1) and the use of unit assessment systems to aggregate and analyse data with a view toward program improvement (Standard 2). Data must indicate that candidates meet professional, state, and institutional standards. Institutions nationally are struggling with meeting these two standards.

The problem, as we see it, is simple and twofold. The obvious problem is that many college faculties prefer not to think about measurement issues, leaving such things to "those folks down the hall" who like to hurl numbers and Greek letters in the air like so many Olympic discus or javelin throwers. One could never hope to achieve their capacity to throw the distance or imagine how they acquired the skills do it! The other problem is related to planning. Most assessment systems have grown like a 600 square foot house with 14 room additions now adding up to 8,000 square feet. Without the benefit of an initial plan, the house may have multiple design flaws, is too big for most folks, and has become an ugly collection of rooms in different styles, built at different times, for a variety of purposes that do not stand together in a useful way. Bathrooms, windows and doors are there, but they may not be where they should be, and there may be too many or too few of them.

So, the problems we face are: (a) assessment anxiety and illiteracy and (b) poor initial planning. The good news, though, is that both are relatively easy to resolve. Start with the second, and you will move more easily to the first. Know what you want, plan it, and tell your friends down the hall the specific data you want. Today's paper may help you do that.

There are also three critical flaws in the typical assessment process, which make meeting psychometric requirements virtually impossible. These flaws result in a hodgepodge or haphazard collection of evidence assembled without use of design frameworks or blueprints (AERA, APA, NCME Section 3, 1999):

- Evidence is typically drawn from a collection of class assignments, designed based on course objectives to determine a course grade and then used as summative assessments of standards to which they may be partially aligned. Thus, they are being used for a purpose for which they were neither designed not intended.

- The collections of artefacts, self-selected by the student or chosen solely on the basis of a tangential relationship to a standard rather than a predetermined alignment with all important aspects of a standard, typically fail to stand the test of construct

representativeness (or domain sampling). Sampling (through a test blueprint) was not the design starting point.

- Decisions made about teacher competency based on a self-assessment through reflection do not stand the test of job-relatedness. In service teachers rarely analyse their work against teaching standards once they are in the classroom full-time, and few schools require reflections to be turned in and reviewed by building administrators.

So, from the outset, some fundamental tenets of establishing validity are jeopardized in the assessment design processes used by most teacher preparation institutions. Validity is not an afterthought. It is a forethought, and **not** with malice. All kidding aside, the secret to achieving validity is to think about what you want to know about teacher candidates in advance, why you need to know it, and what you will do with the information once you get it. In this presentation, we will first describe some planning strategies to control the size and design of the "house," and then we will provide the results of six validity studies of a task-based approach for measuring teacher performance that you can replicate. We will very briefly present a design model for assessment systems that will help institutions get started on the road to validity. We will then present some quick strategies that help provide data on validity and show you what the results look like and how we used them in the decision-making process. We encourage you also to look at some of our other work on creating a psychometric plan can help guide you in your quest for evidence of validity, reliability, and fairness in systematic ways. We have a more in-depth discussion of psychometrics in the workshop materials that are also available to you.

The studies we will share with your today are easy to replicate for other task-based systems and include both judgemental and empirical approaches. They are:

1. Content Validity Study Aligning Tasks with Standards
2. Content Validity Study Based on Expert Judgmental Review
3. Instructional Validity Study of Opportunities to Learn
4. Construct Validity Evidence on Job-Relatedness
5. Content Validity Based on Stakeholder Focus Groups
6. Construct Validity Evidence Based on Empirical Analysis

The tasks we will describe today have been adopted by the Florida Alternative Certification Program and are being used in several Florida colleges of education and in about 45 of the 68 school districts. These tasks have been found to be critical and authentic representations of teachers' required work skills by stakeholders. You will also find one of them in the current paper being prepared by NCATE on K-12 Impact.

Before leaving our brief introduction, we offer pause to ask you to consider the importance of planning. It is not likely that anyone here would support a teaching process that did not begin with an analysis of the curriculum and a detailed set of unit and lesson plans, accompanied by a well-thought out assessment process designed to determine what students learned and how to remediate for those who did not. The same holds true for assessment systems. When we start with a good plan and then look for evidence that we are making good decisions about teachers, we have the data to state with some degree of certainty that we are effective teacher educators and to improve what we do in our teacher preparation programs.

Literature Review

Cureton (1950) concluded that reliability without validity is meaningless. Sadly, most institutions seeking to ensure the psychometric properties of their assessments are relying on inter-rater reliability studies. The statistics produced by analysing scores that show everyone as having met every target produce little of real use. Institutions mistakenly conclude that they have good data instead of recognizing that all they have is consistent data – regardless of what it means in terms of the standards they are attempting to measure. There is no substitute for examining validity and reliability as they are both important to measurement. The tasks that are a part of this system, and the approaches used to validate them, are drawn from a large scale literature review, references for which are provided. A limited version of this presentation of validity evidence has been published in *Practical Assessment, Research, and Evaluation* (Wilkerson and Lang, 2004), and the references from that publication are included in this paper.

Designing Assessment Systems for Valid Results

We have proposed and described a five step design model for creating assessment systems that can provide for valid inferences about teacher candidate competency (Wilkerson and Lang, 2004). The steps of the Competency Assessment Aligned with Teacher Standards (CAATS) model are as follows:

1. Define content, purpose, use, and other contextual factors.
2. Develop a valid sampling plan with frameworks.
3. Create or update tasks aligned with standards and consistent with the sampling plan.
4. Design and implement data tracking and management systems.
5. Choose a psychometric method of analysis and implement it for integrity.

A few comments about aspects of the model that help ensure validity follow:

- In Step 1, designers begin by determining what they want to know (assessment content), why they want to know it (assessment purpose), and what they will do with the information once they get it (assessment use). Each purpose and use are conceptualized and evaluated separately as a matter of validity. We have long known that when data are used for purposes other than those that were intended, validity is sacrificed.

- In Step 2, all relevant standards are identified and aligned into assessment domains, making sampling feasible. A job analysis is conducted to identify critical skills that need to be assessed. Designers guide their analysis by visualizing what each standard looks like in practice when performed by a good teacher, ensuring that all skills deemed essential are assessed. (The pilot has to know how to land the plane!) With that vision in mind, the designer can construct a series of design frameworks, or blueprints, to guide the development of the system, ensuring an effective balance of appropriate assessments along a variety of dimensions, such as:

    - Types of competency (knowledge, skills, dispositions, impact on K-12 learning)
    - Level of measurement inference (high, medium, low)
    - Timing (admission, pre-internship, internship, graduation, post-graduation)
    - Assessment method (tests, products, observed performances, interviews, scales, etc.).

- In Step 3, we ensure that the tasks created or updated are consistent with the sampling plan. It is important to ensure that sight of the plan is not lost and that the alignment remains solid. The specific aspects of the standards assessed in each task should be identified and the language of the standards embedded in the tasks to the extent possible. Once the linkages are made explicit, a content validity study is possible to ensure that the standards are thoroughly addressed and the sampling plan has been followed.

- In Step 4, with data collected and analysed systematically, users have the data needed to conduct additional psychometric studies, all of which can be planned in advance in the psychometric plan called for in Step 5.

- In Step 5, a psychometric plan is developed to ensure that information appropriate to the size and needs of the institution is collected systematically, yielding results that will add to the credibility of the data and the utility, accuracy, and precision of the decisions. Such data is used for decisions about candidates as well as programs.

Planning for Validity in Any Task-Based Assessment System

Using the set of steps outlined above, the task-based system described in this paper was designed with validity in mind, and it is the one we recommend for all assessment systems that lead in one way or another to teacher certification or licensure. If you endorse a transcript for your state, and the state relies to some extent on that endorsement to mean that the teacher has demonstrated a certain level of competence on state and/or national standards, then this series of steps applies. It also applies if you do not play such a role in your state but your conceptual framework commits you to the preparation of well-trained and highly qualified teachers who are likely to perform appropriately in the education profession.

First, as we designed our system, we established that the construct we needed to measure was the teachers' ability to perform on the job – just as we have stated in the previous paragraph. This is a typical construct for any licensure or certification related decision – medical doctor, lawyer, airline pilot, or nail technician. Clearly, what we need to know is if the individuals being trained can fulfil the basic requirements of the job! For example, in the case of teaching, can they plan and deliver a lesson, develop and implement a classroom management plan, determine which students are learning what and remediate accordingly, etc.? If they can keep a great diary, that is a good thing, but do they have to do it on the job and is it as important as knowing their content? Or is it an indirect measure of reflection and continuous improvement that could better be demonstrated through a series of running notes and reflections on the learning of an individual child?

Since there are standards for which teachers and teacher educators are held accountable in Florida, the standards provided a useful tool to identify and design the tasks. Florida teachers have to demonstrate the Florida Educator Accomplished Practices (FEAPs) both as part of their initial certification (regardless of preparation route) and typically as part of their performance appraisals in the districts. These FEAPs were drawn from the national Interstate New Teacher Assessment and Support Consortium (INTASC) Principles, aligning virtually with a one-to-one correspondence, although Florida added two Practices on Ethics and Technology. The tasks listed above are clearly derived from the planning,

communication, assessment, and diversity standards (FEAP 10, 2, 1, 5, 8, 3 and INTASC 7, x, 8, x, and x).

As a matter of design, therefore, the planning issue was to identify and develop those job-related tasks that were critical functions of successful teaching for each FEAP. The process used to do this is described in more detail in the PARE article, referenced above. For now, we note that to plan for valid inferences about teachers, we started with the construct teacher job performance. We operationally defined the content for that construct using the 12 FEAPs, which form the 12 domains from which we sample work. We then visualized the competent teacher performing those standards in order to identify the key job tasks which we analysed using a series of frameworks or matrices – sometimes called blueprints.

Questions as a Design Tool for Credibility (Psychometrics)

Much good research starts with asking questions. The questions we have identified and answered so far from our psychometric plan are listed below. The remaining questions are beyond the scope of this paper but are included in the workshop materials presented at this meeting. We remind our audience that one of the big ideas in psychometrics is that one is never done collecting evidence that the assessments are producing the expected results and not causing any unforeseen problems.

Prerequisite Question:

What construct do we want to measure and for what purpose? The following questions are all based on the answer to that question being "teacher performance for initial certification".

Questions Studied To Date:

1. Does the assessment system provide adequate coverage of the Standards (validity -- content)? -- See Study #1 (Psychometric Plan Question #1).

2. Are the tasks an adequate representation of the job? Are they critical to job performance, authentic, and frequent (validity -- content)? -- See Studies #2, 4 and 5. (Psychometric Plan Question #2)

3. Are procedures in place to ensure that all teachers know the requirements and have adequate opportunity to learn the content and remediate when completion of tasks is not initially successful? (fairness -- EO) -- See Study #3 and 5. (Psychometric Plan Question #13)

4. Do individual tasks fit together (validity -- internal structure)? -- See Study #5 (Psychometric Plan Question #4)

5. Is there evidence that item structure (e.g., proficiency levels o three-point rating scales) is appropriate? (reliability –consistency) -- See Study #6 (Psychometric Plan Question #14)

6. Are scores obtained on teachers sufficiently precise as to have confidence that they could be replicated under different or new conditions or administrations (reliability -- internal consistency and measurement error)? -- See Study #6 (Psychometric Plan Question #10)

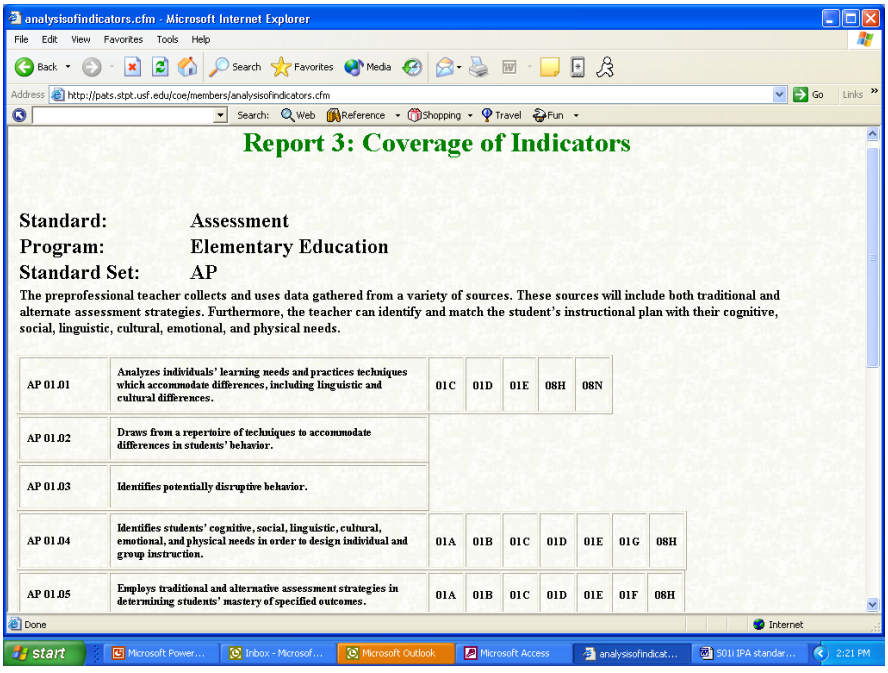
7. Are the data produced useful for decision making about candidates and programs? -- See Study 6. (Psychometric Plan Question #15)

Validity Studies Used for FACP Assessment System of Job-Related Tasks

Study #1: Content Validity Study Aligning Tasks with Standards

By using the FEAP and INTASC principles as the foundation for the definition of minimal competence and task creation, initial evidence of content validity was easy to establish. Aligning tasks to the standards through the creation of charts in a software program we designed made it clear that the tasks were relevant and covered the intent of the standards. We were able to see visually which indicators were used in the design process and which were not, allowing us to determine if we had missed anything important or placed too much focus on any single indicators within the standards. A sample chart for FEAP #1, using the Florida indicators, follows in Figure 1.

Figure 1. Alignment of tasks with FEAP #1 (Content Validity)



In our in-house review of coverage, we noted that multiple tasks tapped the same indicators, but in our professional judgments, this was an appropriate and proportionally correct number of "hits" for some very important indicators. In Florida, the indicators used are considered to be samples of the behaviours. We believed that the repertoire of accommodation techniques (Indicator #2) and the identification of potentially disruptive behaviour (Indicator #3) – the two indicators for which there were no tasks – are important but were adequately assessed through other Standards on learning environment and diversity. We concluded that nothing additional was needed here.

The charts (computer output) themselves served as the report, although discussions such as the one above about the proportionality of coverage and explanations of gaps could

be an appropriate form of documentation to keep if deemed necessary. Since we had other studies in mind, we did not keep these records on file. We were comfortable enough with the tasks to move forward. Our decision at this point was that the study provided adequate support for us to continue the assessment validation and design process.

Study #2: Content Validity Study Based on Expert Judgmental Review

Two former school principals on the design team validated the criticality of each task before it was included in the system, thereby establishing initial evidence of job-relatedness (Wilkerson et al., 2002). A second important validation study was provided through a judgmental review by the staff of a large Florida school district. This was the first stakeholder (or expert) review from outside our College of Education. In this study, prospective users from staff development (outside the Ivory Tower) reviewed the tasks for criticality and authenticity (job-relatedness). Modifications were made based on their feedback to individual tasks, but no tasks were eliminated or added. Feedback was informal, mostly through e-mail and the Florida Department of Education personnel, who relayed suggested changes to us. Again no formal report was written. Where they said, "teachers don't need to do this!" we removed an element of a task.

Records of the conversations and e-mails would be an appropriate form of documentation to keep if deemed necessary. Again, since we had other studies in mind, we did not keep these records on file, many of which were directed to the contractor and then relayed to us second-hand. We were comfortable enough with the tasks to move forward with a field test of the tasks, after making the modifications suggested.

Study #3: Instructional Validity Study of Opportunities to Learn

Once we were confident that the tasks were an adequate representation of the construct and the content, it was time to ensure that teachers would have adequate opportunity to learn the content before being assessed on it. This was the third validity study, often called instructional validity. We note in passing that there are many aspects of validity with many titles. Recent trends in the measurement field, as articulated by Messick (1994) and in the standards which guide this profession (AERA, APA, NCME, 1999) advise us to think about validity as a unitary concept and to collect evidence to support various aspects of it. While the unidimensionality and clarity about the construct measured is critical, the other most important aspect in credentialing is content validity.

In this study, a thorough examination of each task was made to ensure that the target was clear and that opportunities to read and practice the content were provided. This was an in-house study that we conducted, with our first formal report written for the FDOE. Clearly, if teachers do not have the opportunity to learn and remediate, there is a major problem in fairness, and we might exclude people from the profession who have the capacity to teach had we done a better job of teaching them ourselves.

For this kind of study, it is important to look at the details. We checked each criterion on every one of the scoring rubrics for each of the 42 tasks to see if it was aligned with an instructional objective, material in the readings, and a practice activity. We concluded that teachers, for the most part, had adequate opportunity to understand what they were being asked to do in the tasks -- the tasks were aligned with objectives and instruction. A few gaps were identified and filled with additional activities, and the field test continued as planned.

Study #4: Construct Validity Evidence on Job-Relatedness

The fourth validation study (Wilkerson and Lang, 2003a) was conducted during the first year of implementation (2002-2003), after multiple districts had had an opportunity to start using the tasks. One always has to be sure that the construct is appropriately and adequately measured, so we returned again to construct validity, this time in a formal process of evaluation with district HRD users (ACP coordinators and mentors) now familiar with the process and the tasks and able to provide the necessary confirmations.

A survey of district personnel (stakeholders) was administered, using a theoretical framework based on the suggestions of Crocker (1997). In her validity study of the assessments used by the National Board of Professional Teaching Standards (NBPTS), Crocker asked judges to rate the frequency, criticality, and realism of the performance exercises. In the Florida study, frequency and criticality were also used but authenticity was substituted for realism – a term we thought would be more meaningful and current. These three terms were used as an operational definition of "job-relatedness," and they are consistent with our psychometric *Standards* (APA, AERA, and NCME, 1999).

Each of the 42 Assessment tasks was assessed on all three criteria. Each criterion was assessed on a three-point scale as follows:

> Frequency (F): How frequently are the knowledge, skills, and attitudes measured in the tasks evidenced by good teachers in the classroom? How often should they display these skills? 3 = daily or weekly, 2 = monthly, 1 = once a semester , 0 = or not at all.
>
> Criticality (C): How important or critical are the knowledge, skills, and attitudes measured in the tasks? 3 = critically important , 2 = very important , 1 = somewhat important, 0 = not important at all.
>
> Authenticity (A): Determine if teachers really do the kind of work represented in the tasks (or are observed for these behaviors) -- even if they do not typically put the results of their work in a neat folder or write a report/reflection about it. 3 = highly authentic, 2 = moderately authentic, 1 = slightly authentic; 0 = not authentic at all.

Figure 2 shows a segment of what the survey actually looked like.

Figure 2. First three items on Validity Questionnaire to Stakeholder Districts

Task #	Task Name	Frequency	Criticality	Authenticity
01A	Unit Exam/Semester Final	3 2 1 0	3 2 1 0	3 2 1 0
01B	Alternative Assessment	3 2 1 0	3 2 1 0	3 2 1 0
01C	Classroom Assessment System	3 2 1 0	3 2 1 0	3 2 1 0

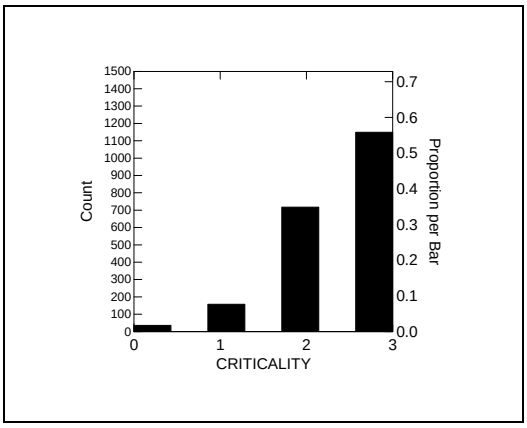
A narrative report was written with simple descriptive statistics and graphic representations of the results. Figures 3 and 4 provide a sample of the data analysis across all tasks with the supporting chart reporting the survey results. (Note similar reports were made by FEAP and task.)

Figure 3. Narrative Discussion of Criticality Results

> "…Across all tasks, more than half of the respondents (56%, n=1149) found the tasks to be critically important and another third (35%, n=717) found them to be very important. Thus, a total of 91% of the responses indicated that these tasks are critical to the performance of a competent teacher. Less than

one-tenth (8%, n=157) found the tasks to be somewhat important, and only 2% (n=35) found them to be unimportant."

Figure 4. Criticality of ACP Tasks



Study #5: Content Validity Based on Stakeholder Focus Groups

In this last of the judgmental studies conducted to date, we assembled representatives from 14 school districts to discuss and respond to a series of questions designed to elicit information again on the criticality and authenticity of tasks but this time adding evidence on the structure of the tasks and the levels of proficiency being used. District personnel were mixed, so that not more than one district person was in each of 5 groups. They were assigned two or three FEAPs and asked to respond to a set of questions. Groups then shared any concerns identified with the larger group.

Stakeholders were asked if the instructions and criteria were clear, if the judgements being made were too harsh or too lenient at the criterion and task levels, if alternative tasks should be provided, and if anything should be added to or eliminated from the system. Results were provided in detail and also summarized. Figure 5 provides an example of the summary report on one question:

Figure 5. Sample Summary Report from Stakeholder Focus Groups

Rubric Level Decisions (cut score): The cut score decision-making process was unanimously supported with only a few suggestions to be harsher rather than more lenient in making the cut between "demonstrated" and "partially demonstrated". There were no suggestions to make the "not demonstrate" decision more lenient. To the contrary, in large group discussion, reviewers felt this would provide too easy an escape mechanism

Study #6: Construct Validity Evidence Based on Empirical Analysis

Early results using the Rasch procedure of Item Response Theory (IRT) with the Florida performance tasks provide empirical evidence of construct validity. The Rasch model allows us to place people and items on the same interval scale, so that we can make predictions about how people of different ability levels are likely to perform on items of different degrees of difficulty. As in classical test theory (CTT), we expect that those who are high achievers are more likely to get hard items correct than are low achievers, etc. The difference here is that we have information that is more useful for individual people and

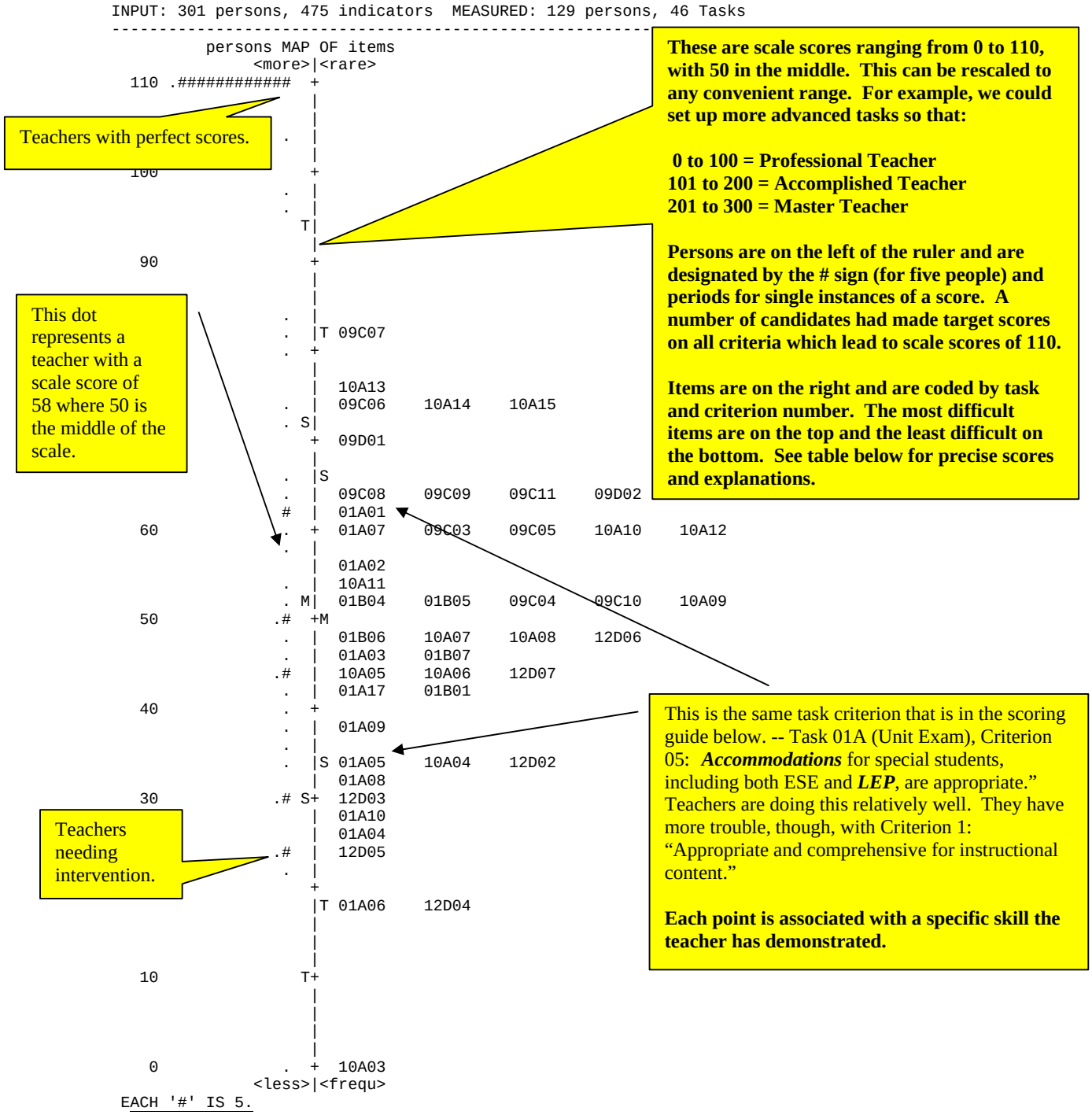
individual items. Hence, we can make conclusions not only about validity in general for all teachers assessed, but we can also hone in on an individual teacher and say that the results are or are not valid for that person! This is a very powerful tool. Maybe the teacher was anxious, or had native language interference, or received poor instruction, or had a rater who was too harsh or erratic, or even cheated on some tasks. IRT helps us identify and help those teachers succeed (or not succeed). A more in depth discussion of that basic concept is beyond the scope of this paper. Suffice it to say here that the successful calibration of items onto an interval level scale (logistic ruler) is an important step for any number of reasons, including future criterion-related studies and other validity, reliability, and fairness studies.

Figure 6 provides the sample logistic ruler, calibrating the items from five of the 42 tasks in the current performance system. Teachers are on the left side, with most getting high scores. Criteria from the rubrics are on the right side and are pretty much normally distributed. The top of the ruler is where we find teachers with high ability and items that are difficult; at the bottom are poorly performing teachers and easy items.

Even at these early calibration stages, where sparse data remain largely unconnected (most teachers have only completed a few tasks), it is possible to confirm that items that were expected to be more difficult are being scaled as more difficult and items expected to be easier are being scaled as less difficult. One example is demonstrated in Figure 6 from task 01A – Unit Exam. Instructors' experience indicates that teachers have less difficulty in making ESE and ESOL accommodations on tests (criterion 5, coded 01A05) than on matching test items to instructional content (criterion 1, coded 01A01). In our state, ESOL is a major focus, and we teach it extensively. Further, the lack of gaps in the scale of items supports the adequacy of coverage of the domains, an indication of content validity.

The ruler in Figure 6 also provides evidence of a simpler and more often overlooked component of validity: operational functionality. An assessment is only as good as the ability to report information that is practical and informative to the user. Percent correct or percentile rank results accompanied by a cut score are weak for these purposes. In the example below, even those who are not statisticians can quickly see that most teachers have mastered the tasks, but that a few are lacking. Outliers among both persons and items are readily observable. Gains on the measures, prerequisite ordering of tasks, gaps and redundancy of items, specific diagnosis of person weaknesses, and the interaction of different tasks are graphically visible. A few points are demonstrated in the callout on Figure 6 to illustrate.

Also in Figure 6, at the bottom, one finds some initial evidence of internal consistency in the reliability statistic reported. The Rasch model generates a version of Cronbach's alpha for us. The separation statistic provided next to it also provides information on our ability to generalize to other items and persons based on the spread of scores.



Data on Items:

```
INPUT: 301 persons, 475 items  MEASURED: 129 persons, 46 items, 3 CATS      3.47
--------------------------------------------------------------------------------
person: REAL SEP.: 2.24  REL.: .83 ... item: REAL SEP.: 1.94  REL.: .79
```

Figure 6. Logistic Ruler (Scale Scores for Items and Persons) on Skill-Based Tasks as Presented to School Districts and College of Education Deans ("marketing tool")

As part of the same analysis, although not demonstrated here, we are able to obtain information on the extent to which certain points on the scale are used adequately to be meaningful. Called thresholds, this helps us identify whether we have too many points and often indicates that three or four points on a scale, or proficiency levels, are all that raters can typically use with consistency. After all, it is tough to tell the difference between a seven and an eight or a four and a five on most tasks!

In deciding whether or not the information is useful, we are able to see that most teachers do very well, but there are a few at the bottom of the scale who should be provided some additional support. Rasch also provides for individual reports that point toward unexpected results where, for example, candidates missed an important area of content that was within their ability range. Finally, at the program level, we are also able to identify some criteria that need to be reviewed to see if we can improve instruction. For example, is there something we need to do about matching test content to instructional outcomes, based on the lower score for item 01A01? These kind of questions and answers help us to see that the assessment system is providing useful data for making decisions – our ultimate purpose in doing all this!

Conclusions

Unfortunately, many college faculty and administrators are afraid of the word "validity." In this paper, we have attempted to show that evidence of validity can be helpful in the following ways:

1. It helps us remember to measure a construct and not just any old thing.
2. It gives us a good reason to reduce what we measure to a set of meaningful and useful tasks.
3. It keeps the focus on decision-making.
4. It provides a valuable role for involving stakeholders.
5. It gives us evidence that what we are doing is useful.
6. It helps us diagnose areas for improvement in both candidates and programs.
7. It helps us guard against making bad decisions that adversely affect children and protected populations.
8. It helps us remember to check that we teach what we test.
9. It helps us focus on collecting data in ways that can yield studies important for research purposes.
10. It makes room for those who like words and those who like numbers.

Those are just a few of the benefits and perks. Above all, we hope you will remember to always work from a plan and never stop collecting data on effectiveness. That is what we tell our graduates to do, and we need to model it ourselves.

References

American Educational Research Association, American Psychological Association, and National Council of Measurement in Education (1999). *Standards for educational and psychological testing.*

Barrett, H. (2004, March). Differentiating electronic portfolio systems and online assessment management systems. Paper presented at the Annual Meeting of the Society for International Technology in Education (SITE), Atlanta, Georgia.

Bond, T. G. & Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences,* Mahwah, NJ, Lawrence Erlbaum.

Bohlig M., Fisher, W.P. Jr., Masters, & G.N., Bond, T. (1998) Content Validity and Misfitting Items. *Rasch Measurement Transactions, 12:1,* 607

Briggs, D.C. & Wilson, M. (2003). An introduction to multidimensional measurement using Rasch models. *Journal of Applied Measurement, 4:1,* 87-100.

Council of Chief State School Officers. (1998). *Key state education policies in K-12 education: Standards, graduation, assessment, teacher licensure, time, and attendance: A 50-state report*. Washington, D.C.: Author.

Crocker, L. (1997). Assessing content representatives of performance assessment exercises. *Applied Measurement in Education. 10:1,* 83-95.

Cureton, E. E. (1950). Validity, Reliability, and Baloney, *Educational and Psychological Measurement, 10,* 94-96.

Darling-Hammond, L., Chung, R., & Frelow, F. (2002). Variation in teacher preparation: How well do different pathways prepare teachers to teach?. *Journal of Teacher Education. 53:4,* 286-302.

Engelhard, G. (1994). Resolving the attenuation paradox. *Rasch Measurement Transactions, 8:3,* 379.

Fisher, W.P. (2001). Invariant thinking vs. invariant measurement. *Rasch Measurement Transactions 14:*4, 778-81.

Hawley, W.D. (1985). Designing and implementing performance-based career ladder plans. *Educational Leadership. 43:3,* 57-61.

Impara, J. C. & Plake, B. S. (2000). *A comparison of cut scores using multiple standard setting methods*. Paper presented at the Large Scale Assessment Conference, Snowbird, UT.

Ingersoll, G. M. & Scannell, D. P. (2002). *Performance-based teacher certification: creating a comprehensive unit assessment system.* Fulcrum, Golden, CO.

Lee, W.W. & Owens, D. L. (2001). Court Rulings Favor Performance Measures. *Performance Improvement, 40:4, 35-40.*

Linacre J.M. (1996) True-Score Reliability or Rasch Statistical Validity? Rasch Measurement Transaction 9:4 p. 455-6

Linacre, J.M. (2003) A User's Guide to Winsteps Rasch-Model Computer Programs, Chicago.

Lissitz, R. and Huynh, H. (2003). Vertical equating for state assessments: Issues and solutions in determination of adequate yearly progress and school accountability. *Practical Assessment, Research & Evaluation*. *8:1*. Retrieved December 26, 2003 from http://PAREonline.net/getvn.asp?v=8&n=10.

Lopez, W.A. (1996). Communication validity and rating scales. *Rasch Measurement Transaction. 10:1*, 48.

Mehrens, W. A. (1991). Using Performance for Accountability Purposes: Some Problmes. *ERIC Document Reproduction Service, ED333008.*

Myford, C. M. & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement.* 4:4, 386-421.

National Commission on Teaching and America's Future (2003). No Dream Denied: A Pledge to America's Children. New York: Author.

National Commission on Teaching and America's Future (1996). What matters most: Teaching for America's future. New York: Author.

National Council for Accreditation of Teacher Education (2001). *Professional Standards for the Accreditation of Schools, Colleges, and Departments of Education.* Washington, D.C.: Author.

National Research Council (U.S.), Committee on Assessment and Teacher Quality (2001). *Testing teacher candidates: The role of licensure tests in improving teaching quality.* Committee on Assessment and Teacher Quality, Center for Education, Board on Testing and Assessment, Division on Behavioral and Social Sciences and Education, National Research Council, Mitchel, K.J., Robinson, D.Z., Plake, B.S., Knowles, K.T., editors. Washington, DC: National Academy Press.

Nweke, W. & Noland, J. (1996). *Diversity in Teacher Assessment: What's Working, What's Not?* Paper presented at the Annual Meeting of the American Association of Colleges for Teacher Education, Chicago, Ill. *ERIC Document Reproduction Service, ED393828.*

Pascoe, D. & Halpin, G. (2001). Legal issues to be considered when testing teachers for initial licensing. Paper presented at the Annual Meeting of the Mid-South Educational Research Association, Little Rock, AK. *ERIC Document Reproduction Service, ED460162.*

Rebell, M.A. (1991). Teacher Performance Assessment: The Changing State of the Law. *Journal of Personnel Evaluation in Education, 5*, 227-235.

Rudner, L. M. (2001). *Measurement Decision Theory.* Retrieved May 28, 2004: http://edres.org/mdt/home3.asp

Southeast Center for Teaching Quality (2003a). NCLB Teaching Quality Mandates: Findings and Themes from the Field, *Best Practices and Policies. 3: 4*, December. Chapel Hill, NC: Author.

Southeast Center for Teaching Quality (2003b). How do teacher learn to teach effectively? Quality indicators from quality schools. *Best Practices and Policies. 2: 7*. January. Chapel Hill, NC: Author.

Southeast Center for Teaching Quality (2003c). Performance-based teacher compensation: Learning from the lessons of history. *Best Practices and Policies. 3:1*, May. Chapel Hill: Author.

Trochim, W. M. (2002). *The Research Methods Knowledge Base, 2nd Edition*. Available online at: http://trochim.human.cornell.edu/kb/index.htm.

Wilkerson, J.R. & Lang, W.S. (2004b). A standards-driven, task-based assessment approach for teacher credentialing with potential for college accreditation. *Practical Assessment, Research & Evaluation*, 9(12). Retrieved June 20, 2004 from http://PAREonline.net/getvn.asp?v=9&n=12 .

Wilkerson, J.R. & Lang, W.S. (2003a). Florida Alternative Certification Program Assessment System: *Analysis of District Coordinators' Validity Questionnaire on Assessment Tasks.* Report: Bureau of Teacher Certification, Florida Department of Education, Tallahassee, FL.

Wilkerson, J.R., & Lang, W.S. (2003b, December 3). Portfolios, the Pied Piper of teacher certification assessments: Legal and psychometric issues. *Education Policy Analysis Archives, 11:45*. Retrieved December 20, 2003 from http://epaa.asu.edu/epaa/v11n45/.

Wilkerson, J., Lang, W.S., Hewitt, M., Egley, R., & Stoddard, K. (2002). *Florida Alternative Certification Program Assessment System.* Bureau of Teacher Certification, Florida Department of Education, Tallahassee, FL.

Wright, B.D., & Stone, M.H. (2004). *Making Measures*. Chicago, The Phaneron Press.

Wright, B. D. & N. A. Panchapakesan (1969). A procedure for sample free item analysis. *Educational and Psychological Measurement, 29*, 23-48.

Zirkel, P. (2000). Tests on Trial. *Phi Delta Kappan. 8: 10,* 793-794.